One-Sided Group Sequential Procedures for Clinical Trials

by

Chih-Hsiang Ho
University of Minnesota

Technical Report No. 462
January 1986

One-Sided Group Sequential Procedures for Clinical Trials

Abstract

I propose an asymmetric stopping rule which allows the experimenter to terminate a clinical trial early for a sufficiently negative result and to continue to a specified number of patients otherwise. If interim data are positive we are willing to wait in order to measure the various safety factors. A sufficiently negative interim result will terminate the trial to either (1) minimize exposure to an apparently inferior treatment, or (2) cut the losses of a pharmaceutical company whose product is inferior.

Key Words: Clinical trials; Repeated significance tests; Group sequential designs; Interim analyses; Nominal significance levels; Stopping rules.

1. INTRODUCTION

In many clinical trials that compare two treatments on the basis of short-term measures of effectiveness, the data are reviewed periodically to identify any trends which may suggest early termination. One possibility is the repeated application of standard statistical tests with the usual critical values. However, when the null hypothesis is true, the probability of obtaining a "significant" outcome at least once during a sequence of such tests is substantially higher than the nominal significance level (see Armitage, Mcpherson & Rowe (1969)). Pocock (1977) suggests a larger constant critical value for each test, the maximum number of such tests being specified in advance. O'Brien and Fleming (1979) suggest being very conservative early by using large critical values and decreasing the nominal critical value during the course of the trial so that the actual level is close to the nominal level at the test scheduled at the end of the trial. Lan and DeMets (1983) have shown how to use these methods without having to specify the number of tests in advance but only the form of boundary. Demets and Ware (1980) also consider three possible modifications of the group sequential method for sequential monitoring of clinical trials testing a one-sided hypothesis. Another approach which incorporates a "range of equivalence" into a formal stopping rule for the trial using an extension of the group sequential design has been recently discussed by Freedman, Lowe and Macaskill (1984).

There are many trials in which the investigator(s) would not be willing to stop the trial on the basis of superior efficacy of the treatment. For this would result in limited information regarding safety considerations (as well as for secondary efficacy measurements). For example, few if any trials sponsored by pharmaceutical companies would be stopped on the basis of favorable efficacy. Regulatory agencies require that the drug be used on a reasonably large number of patients so that serious adverse experiences with moderately low incidence rates can be detected.

I propose a design which allows the drug developer to terminate the development program early if the new drug is not efficacious and to continue the trial to a specified number of patients as long as the results are positive. So if accumulating data show a benefit, the trial (program) will continue so that information concerning the various safety factors will accrue. But if interim data are sufficiently negative then the trial will end early, thus minimizing ineffective treatment and saving resources that can be better allocated elsewhere. Clearly, the objectives of the one-sided stopping rules proposed here are quite different from those suggested by Demets and Ware (1980). The major question that is addressed in this paper is: how negative need the results be before the trial should be stopped?

2. PROPOSED DESIGN

Suppose that a clinical trial is performed to compare a new treatment T to a control C with respect to mean outcome on a continuous response variable. Assume that the treatment response variable $X_T \sim N(\mu_T, \sigma^2)$ and the control variable is $X_C \sim N(\mu_C, \sigma^2)$. Following the group sequential test of Pocock (1977), the data are tested after every equally divided group of 2n patients is entered, up to a maximum of N groups of subjects. Assume σ^2 is known. This assumption is not very restrictive when n is moderately large.

Let $\mu_T - \mu_C = \delta$. Consider testing

$$H_0 : \delta \leq 0 \text{ vs. } H_A : \delta > 0.$$

At the mth stage, we shall use as test statistic,

$$\bar{d}_m = \sum_{j=1}^{m} (\bar{t}_j - \bar{c}_j)/m,$$

where $\bar{t}_j$ and $\bar{c}_j$ represent the observed mean responses for treatments T and C, respectively, in the jth subsample or "group". That is,

$$\bar{t}_j = \sum_{k=1}^{n} t_{jk}/n \qquad \text{and} \qquad \bar{c}_j = \sum_{k=1}^{n} c_{jk}/n$$

where $t_{j1}, \ldots, t_{jn}$ are the n responses of jth group with treatment T and the c_{jk}'s are those with treatment C. For m = 1, ..., N, define

$$Z_m = \bar{d}_m / \sqrt{(2\sigma^2/mn)}.$$

Consider the following testing procedure:

> Stop the trial at stage m and decide that T is not effective (i.e., accept H_0) if $Z_m \leq b_m$ $(m = 1, \ldots, N-1)$. If $Z_m > b_m$ then take the next subsample and repeat the process. If testing reaches the Nth stage, reject H_0 if $Z_N > b_N$ and accept otherwise.

The sequence $\{b_1, \ldots, b_N\}$ is chosen to have appropriate total rejection probabilities.

3. CONSTANT NOMINAL SIGNIFICANCE LEVEL

A simple version of the boundary defined in Section 2 is given by

$$b_N = -b_{N-1} = \ldots = -b_1.$$

A design with this boundary has a constant nominal significance level. For given overall significance level, it can be difficult to calculate b_N. It is relatively straightforward, however, to find a bound for b_N. One is given in the next theorem. The theorem says that a conservative procedure is to ignore the interim analyses!

Theorem 1 If the overall type I error is α then $b_N \leq z_\alpha$ for all α, where $\Phi(z_\alpha) = 1 - \alpha$.

Proof:

$$\alpha = \Pr\{Z_m > b_m,\ m = 1, \ldots, N \mid \delta = 0\}$$

$$= \Pr\{Z_N > b_N \mid \delta=0\}\Pr\{Z_m > b_m,\ m = 1, \ldots, N-1 \mid Z_N > b_N,\ \delta=0\}$$

$$= [1 - \Phi(b_N)]\Pr\{Z_m > b_m,\ m = 1, \ldots, N-1 \mid Z_N > b_N,\ \delta=0\}$$

$$\Longrightarrow \Phi(b_N) \leq \Phi(z_\alpha) \Longrightarrow b_N \leq z_\alpha. \quad \square$$

For a two-sided symmetric test as discussed in Section 1, if one continually tests accumulating data for "significant results", one could eventually reach nominal significance by chance alone. Using the same conventional critical value at each test will raise the type I error beyond the nominal level associated with the critical value. The degree of this increase depends in part on the number of tests. For example, suppose one uses a conventional ±1.96 critical value associated with a 5 percent significance level for a single test. After five tests the type I error would be .142 (see Armitage et al., 1969) and would be considered unacceptable. Thus, some adjustment is necessary for the symmetric boundaries described in Section 1. Theorem 1 says that proceeding with no adjustment is conservative in the one-sided case.

To calculate exact sizes of the tests I used numerical methods based on Armitage et al. (1969) and Pocock (1977). Table 1 shows the actual type I errors if the conventional critical values were used. For the tabled values the differences between the actual and nominal levels with $\alpha' = .01$ or $.05$ are negligible. That is, for

the one-sided test proposed here, the conclusions of the trials are almost independent of the number of interim analyses performed if α is small and N is not too large. This is quite different from the two-sided symmetric setting in which the conclusions are markedly dependent on the stopping rule.

TABLE 1

The actual size α for the test with $b_N=z_{\alpha'}$, the conventional critical value corresponding to α'

N	α'	b_N	α
2	.01	2.3263	.0100
	.05	1.6449	.0499
	.1	1.2816	.0999
3	.01	2.3263	.0100
	.05	1.6449	.0499
	.1	1.2816	.0996
4	.01	2.3263	.0100
	.05	1.6449	.0499
	.1	1.2816	.0990
5	.01	2.3263	.0100
	.05	1.6449	.0499
	.1	1.2816	.0986

4. GROUP SIZE DETERMINATION

The group size 2n is independent of the size of the test and thus the calculation of the boundaries. However, the group size plays an important role when the power of the test is considered.

Following Pocock (1977), let $n = 2(\Delta\sigma/\delta)^2$ (i.e. $\Delta = \delta\sqrt{n}/\sigma\sqrt{2}$), so that it is convenient to express δ in terms of σ. Tables 2 to 5 show the suitable value of n given N, α, δ and power π. The value of n will need to be rounded off resulting in a slight change in π; this difference should be negligible in practice. As an example of these tables suppose $\alpha = .05$, $\delta = .5\sigma$ and $\pi = .9$. Then for N = 3 groups, $\Delta = 1.69$ from Table 4, so that the required number of patients per group is $2n = 4\times(1.69/0.5)^2 = 45.7$ and the maximum number of patients for the whole trial is $2nN = 46\times3 = 138$. This is not exact since n has been rounded to an integer, but the error is slight. Another interesting result is suppose $\alpha = .01$ and $\pi = .95$, then Tables 2 to 5 say that the required number 2n of patients per group is 63.1, 31.5, 21.0, 15.8 for N = 1 to N = 4 respectively. Hence, if we multiply 2n by N correspondingly this gives approximately the same value: 63.1. In other words, if one fixes α and π then the subgroup size 2n for a design with N = k is approximately equal to the total sample size corresponding to a one-sided fixed sample test (i.e., N = 1) divided by k.

5. THE ROLE OF α

The appropriate design of a trial depends among other factors on its purpose, the availability of patients, and costs. In practice one may be forced to fix at least one of the following quantities: n, N or 2nN. The next theorem, which is proved by Ho

(1986, p. 34), says that the average number of patients (ASN) treated before termination of the trial is smaller when α is larger.

Theorem 2 Given n, N and δ, the ASN is a monotonically decreasing function of α.

The numerical results in Figure 1 illustrate Theorem 2 for a fixed total number of patients $2nN = 120$. There are three different designs corresponding to $N = 1$, 2, and 3. The standard deviation σ is assumed to be known and δ ranges from -1σ to $+1\sigma$ with increment $.1\sigma$. It shows that stopping tends to be especially early when δ is small. For example, if $\alpha = .25$ and $\delta = -1\sigma$ then the trial will terminate with probability nearly one at the first interim analysis because of a negative result.

Figure 1 also suggests that given α, δ and 2nN, the ASN is a monotonically decreasing function of N. This means that there are savings in terms of sample size to be gained by having more interim analyses. I have not been able to prove this intuitive result; the changes in the probabilities of early termination when N and n change are not as clear as they are in Theorem 2. I also calculated the power of the tests for the cases considered in Figure 1. Table 6 shows that the power of the test depends on δ and on the size α, but it is not very sensitive to changes in N if 2nN is fixed. The power is exactly the same up to 3 decimal places for N =

1, 2, and 3.

TABLE 6

Values of power given 2nN=120, N=1 or 2 or 3

		α		
		.01	.05	.1
	.1	.038	.136	.232
	.2	.109	.291	.426
	.3	.247	.499	.641
	.4	.446	.708	.818
δ/σ	.5	.660	.863	.928
	.6	.832	.950	.978
	.7	.934	.986	.995
	.8	.980	.997	.999
	.9	.995	1.000	1.000

In summary, lower levels of α provide flexibility in choosing N but trials with higher α levels have fewer patients on average and have higher power. In choosing N, one has to compromise between the group sample size 2n and the power π based on Tables 2 to 5.

6. CONCLUSION

The one-sided group sequential boundaries proposed here are designed for the possibility of stopping a trial early on the basis of negative results. It is not proposed as a competitor to the procedures proposed previously in the literature since their objectives are quite different. For the reasons discussed, the gain of periodic review, stopping only for negative results, is obtained with virtually no loss in inferential ability.

Acknowledgments

I would like to thank my thesis advisor, Professor Donald A. Berry, for his help throughout this work. In addition, a portion of the research was facilitated by a grant from the University of Minnesota Computer Center.

References

Armitage, P.; McPherson, C.K. and Rowe, B.C. (1969). "Repeated significance tests on accumulating data," J. Roy. Statist. Soc. Ser. A, 132, 235–244.

Demets, D.L. and Ware, J.H. (1980). "Group sequential methods for clinical trials with a one–sided hypothesis," Biometrika, 67, 651–660.

Freedman, L.S.; Lowe, D. and Macaskill, P. (1984). "Stopping rules for clinical trials incorporating clinical opinion," Biometrics, 40, 575–586.

Ho, C.H. (1986). "One–sided sequential stopping boundaries for clinical trials: classical and Bayesian approaches," Ph.D. thesis, University of Minnesota, School of Statistics.

Lan, K.K.G. and Demets, D.L. (1983). "Discrete sequential boundaries for clinical trials," Biometrika, 70, 659–663.

O'Brien, P.C. and Fleming, T.R. (1979). "A multiple testing procedure for clinical trials," Biometrics, 35, 549–556.

Pocock, S.J. (1977). "Group sequential methods in the design and analysis of clinical trials," Biometrika, 64, 191–199.

TABLE 2

Values of Δ and required number of patients per group with N = 1 (fixed sample size)

α	b_1	π	Δ	$2n^*$
.01	2.326	.5	2.32	21.6
		.75	3.00	36.0
		.9	3.61	52.1
		.95	3.97	63.1
.02	2.054	.5	2.05	16.9
		.75	2.73	30.0
		.9	3.34	44.5
		.95	3.70	54.7
.03	1.881	.5	1.88	14.2
		.75	2.56	26.1
		.9	3.16	40.0
		.95	3.53	49.7
.04	1.751	.5	1.75	12.3
		.75	2.43	23.5
		.9	3.03	36.8
		.95	3.40	46.1
.05	1.645	.5	1.65	10.8
		.75	2.32	21.5
		.9	2.93	34.3
		.95	3.29	43.3
.1	1.282	.5	1.28	6.6
		.75	1.96	15.3
		.9	2.56	26.3
		.95	2.93	34.3
.15	1.036	.5	1.04	4.3
		.75	1.71	11.7
		.9	2.32	21.5
		.95	2.68	28.8
.2	.842	.5	.84	2.8
		.75	1.52	9.2
		.9	2.12	18.0
		.95	2.49	24.7
.25	.675	.5	.68	1.8
		.75	1.35	7.3
		.9	1.96	15.3
		.95	2.32	21.5

* Multiply each value by σ^2/δ^2

TABLE 3

Values of Δ and required number of patients per group; for N = 2 and various values of power π

α	b_2	π	Δ	$2n^*$
.01	2.326	.5	1.65	10.8
		.75	2.12	18.0
		.9	2.55	26.0
		.95	2.81	31.5
.02	2.054	.5	1.45	8.4
		.75	1.93	14.9
		.9	2.36	22.2
		.95	2.62	27.4
.03	1.881	.5	1.33	7.1
		.75	1.81	13.1
		.9	2.24	20.0
		.95	2.49	24.9
.04	1.751	.5	1.24	6.1
		475	1.72	11.8
		.9	2.14	18.4
		.95	2.40	23.1
.05	1.645	.5	1.16	5.4
		.75	1.64	10.8
		.9	2.07	17.1
		.95	2.33	21.6
.1	1.281	.5	.91	3.3
		.75	1.38	7.7
		.9	1.81	13.2
		.95	2.07	17.1
.15	1.035	.5	.73	2.2
		.75	1.21	5.9
		.9	1.64	10.8
		.95	1.90	14.4
.2	.836	.5	.60	1.4
		.75	1.07	4.6
		.9	1.51	9.1
		.95	1.76	12.5
.25	.657	.5	.48	.9
		.75	.96	3.7
		.9	1.40	7.8
		.95	1.66	11.0

* Multiply each value by σ^2/δ^2

TABLE 4

Values of Δ and required number of patients per group; for N = 3 and various values of power π

α	b_3	π	Δ	$2n^*$
.01	2.326	.5	1.34	7.2
		.75	1.73	12.0
		.9	2.08	17.4
		.95	2.29	21.0
.02	2.054	.5	1.19	5.6
		.75	1.58	9.9
		.9	1.93	14.8
		.95	2.14	18.2
.03	1.881	.5	1.09	4.7
		.75	1.48	8.7
		.9	1.83	13.3
		.95	2.04	16.6
.04	1.751	.5	1.01	4.1
		.75	1.40	7.8
		.9	1.75	12.3
		.95	1.96	15.4
.05	1.645	.5	.95	3.6
		.75	1.34	7.2
		.9	1.69	11.4
		.95	1.90	14.4
.1	1.280	.5	.74	2.2
		.75	1.13	5.1
		.9	1.48	8.8
		.95	1.69	11.5
.15	1.029	.5	.6	1.4
		.75	.99	4.0
		.9	1.35	7.2
		.95	1.56	9.7
.2	.820	.5	.49	1.0
		.75	.89	3.1
		.9	1.25	6.2
		.95	1.46	8.6
.25	.621	.5	.40	.6
		.75	.80	2.6
		.9	1.18	5.5
		.95	1.41	7.9

* Multiply each value by σ^2/δ^2

TABLE 5

Values of Δ and required number of patients per group; for N = 4 and various values of power π

α	b_4	π	Δ	$2n^*$
.01	2.326	.5	1.16	5.4
		.75	1.50	9.0
		.9	1.81	13.1
		.95	1.99	15.8
.02	2.054	.5	1.03	4.2
		.75	1.36	7.4
		.9	1.67	11.1
		.95	1.85	13.7
.03	1.881	.5	.94	3.5
		.75	1.28	6.5
		.9	1.58	10.0
		.95	1.76	12.4
.04	1.751	.5	.88	3.1
		.75	1.21	5.9
		.9	1.52	9.2
		.95	1.70	11.5
.05	1.644	.5	.82	2.7
		.75	1.16	5.4
		.9	1.47	8.6
		.95	1.65	10.9
.1	1.277	.5	.64	1.7
		.75	.98	3.8
		.9	1.29	6.6
		.95	1.47	8.6
.15	1.021	.5	.52	1.1
		.75	.86	3.0
		.9	1.17	5.5
		.95	1.36	7.4
.2	.801	.5	.43	.7
		.75	.78	2.4
		.9	1.10	4.8
		.95	1.30	6.8
.25	.577	.5	.35	.5
		.75	.72	2.1
		.9	1.07	4.6
		.95	1.30	6.7

* Multiply each value by σ^2/δ^2

FIGURE 1

ASN vs. δ/σ when $2nN = 120$

1: N=2, α=.01; 2: N=2, α=.05; 3: N=2, α=.25;

4: N=3, α=.01; 5: N=3, α=.05; 6: N=3, α=.25.